Paper Title: CITADEL: A Semi-Supervised Active Learning Framework for Malware Detection Under Continuous Distribution Drift

Md Ahsanul Haque[1], Ismail Hossain[1], Suresh Kumar Amalapuram[2], Vladik Kreinovich[1], and Mohammad Saidur Rahman[1]
[1]University of Texas at El Paso, El Paso, TX, USA
[2]Indian Institute of Technology Hyderabad, Hyderabad, India

Abstract: Android malware evolves rapidly, leading to concept drift that degrades the performance of traditional machine learning (ML)-based detection systems. While recent approaches incorporate active learning and hierarchical contrastive loss to handle this drift, they remain fully supervised, computationally expensive, and perform poorly on real-world datasets with long temporal spans. Our evaluation highlights these limitations, particularly on LAMDA, a 12-year longitudinal Android malware dataset exhibiting substantial distributional shifts. Moreover, manual expert labeling cannot scale with the daily emergence of over 450,000 new malware samples, leaving most samples unlabeled and underutilized.

To address these challenges, we propose CITADEL, a robust semi-supervised active learning framework for Android malware detection. To bridge the gap between image-domain semi-supervised learning and binary feature representations of malware, we introduce malware-specific augmentations, Bernoulli bit flips and masking, that simulate realistic drift behaviors. CITADEL further integrates supervised contrastive loss to improve boundary sample discrimination and combines it with a multi-criteria active learning strategy based on prediction confidence, L_p-norm distance, and margin uncertainty. This enables efficient adaptation under limited labeling budgets. CITADEL is computationally efficient, achieving $24\times$ faster training and $13\times$ fewer operations compared to prior methods. Extensive evaluation on three large-scale Android malware benchmarks, APIGraph, Chen-AZ, and LAMDA, demonstrates that CITADEL outperforms prior work, achieving F1 gains of 1%, 7%, and 34%, respectively, using only 40% labeled samples.